Modelling Proteins Conformation in Solution. Part I: A Parallel GA Engine for Protein Conformational Space Mapping

Carlos A. Del Carpio [1]
carlos@translell.tutkie.tut.ac.jp

Valentin Gogonea [2]
gogonea@organik.uni-erlangen.de

[1] Lab. for Informatics & AI in Molecular and Biological Sciences
Dept. of Ecological Engineering, Toyohashi University of Technology
Tempaku, Toyohashi 441. Japan

[2] Computer Chemistry Center, Institute for Organic Chemistry
Erlangen-Nurnberg University, Nagelsbachstr. 25, D-91052 Erlangen, Germany

Abstract

This article is the first of a series of papers describing the development of an automatic system for prediction of the three dimensional conformation of proteins in solution. In this first part we discuss the implementation of the protein conformational space mapping engine. This is a procedure based on a robust parallel genetic algorithm which runs on a network of transputers. We describe aspects of the algorithm related to the major factors that influence the protein folding process and describe their implementation within the scheme of the evolutionary algorithm. Among them, we make a throughout review of the co-operativity of emergent partial secondary structures as the evolutionary process proceeds and its effects on the stability of new generated conformers as well as a better performance of the GA. We then undertake the hydrogen bond and synthesize the demographic trends in known proteins suggested by Stickle et. al., and also implement them as an index of goodness assessment of the generations of protein conformers. Finally, we make an intensive analysis of the packing of the amino acid side chains and show how a hybrid algorithm can utter a relaxation of the perturbations brought about by the operations of the GA, and the genuine improvement of the overall process. In the second paper of this series we propose guidelines under which we implement the solvent effect which in concourse with the above mentioned factors results in a system for protein 3D structure prediction in solution.

1 Introduction

One of the most challenging problems in molecular biology nowadays is the prediction of the three dimensional structures of biomolecules such as proteins from the primary amino acid sequence. Many attempts have been done to tackle the problem [1, 2, 3], which span from data analysis and its application to homology based protein design to more sophisticated methods based on derivation of potential functions for protein structure determination. The problem is further complicated when attempts are made in order to predict the structure of the protein within its natural environment, that is in a water solution, mixed with other components of a biological system.

Here we present our first paper of a series describing our theoretical approach for the development of a comprehensive model for the prediction of protein structure in a liquid environment. In this first paper we describe the methodology for exploring the potential energy hypersurface of the protein in gas phase or solution: the Genetic Algorithm (GA), and discuss practical aspects of its implementation in a transputer network. In the second paper we introduce a solvent force field to account for solvent-protein interaction in solution. Our solvent force field is made up of two interaction energy terms: (a) a hydrophobic interaction term which is calculated as a function of solvent-accessible surface area of the protein and (b) an electrostatic interaction energy term which is evaluated by our modification of the generalized Born equation.

Recently, we proposed an evolutionary algorithm for the prediction of polypeptide three dimensional structure [4, 5, 6]. The features of the algorithm are (a) each "chromosome" (protein conformer) is coded as a list of rectangular coordinates of the atoms composing the amino acids from the primary sequence, and (b)the fitness function introduced into the GA as a means for the evaluation of each chromosome and to decide on its "survival", "reproduction", or "mutation". This fitness function takes into account several factors which are considered to contribute largely in the process of protein folding [4, 5, 6]. In addition we describe here three new factors. (a) The co-operativity among protein internal partial structures. Co-operative interaction is especially observed among α helices within the protein and its effects are an increase in stability of the protein and consequently reduction of hundreds of thousands of partially folded intermediates. (b) The second aspect is related with the number of hydrogen bonds allowed by a particular conformer. Specific "demographic" trends identified in proteins, whose three-dimensional structure was determined by X-ray analysis, show that there is a specific number of hydrogen bonds that a protein may have as function of its conformer's helix and sheet content. The comparison of theoretically determined number of hydrogen bonds with that obtained from demographic studies can be used to asses the efficiency of a determined pathway of protein folding. (c) The third aspect discussed here refers to the conformation of the side-chains residues within the different protein backbones generated by the GA engine.Standard criteria for bonding among pairs of amino-acids at both the backbone and the side chain level are extracted from crystallographic analysis of proteins already known. This leads to the optimization of side-chain. The optimization module uses a repulsion potential to resolve the overlapping of non-bonded atoms belonging to different residues and a function which take into account the packing and compactness of the molecule based on an atomic contact potential and the steric energy of the molecule. Thus, we obtained a hybrid GA conformational space searching algorithm by adding this optimization procedure to the conventional GA paradigm.

We illustrate the results of these improvements on the original GA, using the protein CRAM-

Figure 1: Representation of polypeptide conformers in the GA engine

Nature, however selects one of this huge number of conformations which is the one that accomplishes the biological role. The problem in protein engineering is the prediction of the three dimensional structure of this active conformer, and elucidate the structure-function relationship. In order to tackle with these two aspects of protein engineering, we have introduced a new algorithm for mapping the conformational space of a protein. Our algorithm implements an artificial evolutionary process which mimics natural evolution, firstly used by John Holland

and his colleagues [7] and currently known as the Genetic Algorithm. The details of the GA paradigm abound in the literature[8, 9, 10].

The algorithm starts with the randomic generation of the first population of conformers. Then the fitness evaluation of each individual is performed applying the three main operators of the GA, i.e. the selection, crossover and mutation to produce the next generation of protein conformers. The initial population of chromosomes in our GA is created by generating strands of linked monoacid structures following the order of the given primary sequence [4] as illus trated in Fig.1. Each amino acid is extracted from a file containing all 20 geometry-optimized amino acid structures.

As in natural selection, in the artificial paradigm of evolution, conformers of one population remain into the next generation if "they are suited to the environment". In the case of the artificial GA for our problem this consists in examining whether the conformer complies with the internal energy constrains as well as with the restrains due to the interaction of the molecule with its environments. The evaluation of the goodness of a certain conformer is carried out at several levels. The first level is the intramolecular atomic interaction. Here we use a molecular mechanics force field to calculate the internal potential energy of the protein[7]. The second level of evaluation of the fitness of a protein conformer is given by the constrains resulting from interaction of protein subunits on one hand and the interaction of the protein with its environments on the other.

The latter aspect is treated in a separate article which constitutes the second part of this paper. The subunit interactions within the protein which are not accounted for by the molecular force field method, are those concerned with the interactions among parts of protein chain with distinct characteristics (i.e. α-helices, β-sheets)that begin to form as the evolutive process is going on. Similarly, the hydrogen bonding, which originated as an electrostatic by-product of the interatomic interactions, is not properly taken into account by the algorithm. Hence, results of crystallographic statistics of hydrogen bonds formed in already known structures are incorporated into the scheme of fitness evaluation.

2.1 Co-operativity in the Process of Protein Folding

The behavior of proteins at folding is highly co-operative. This observation made by Freire et. al.[11] is supported by the fact that intermediate conformations of the protein between the folded and unfolded states lack thermodynamic stability. The co-operative behavior is a peculiarity of its structural elements which interact locally. Moreover, a higher order co-operative interaction appears between folding units which are formed in later stages of the folding process. Restriction of solvent accessibility to non polar residues located in regions complementary to those that can undergo alternate folding paths may constitute the major contribution to the co-operative folding behavior of proteins. This entropically uncompensated and energetically unfavorable solvent exposures may characterize most of the partially folded states but not the completely folded state, thus minimizing the population of partially folded intermediates throughout a folding processes. Nature certainly selects the protein folding path where only entropically favorable processes occur. This natural behavior of the system reduces to negligible amounts the probability of hundreds or thousands of partially folded intermediates with a potential to become populated during a folding process. We have adopted the statistical thermodynamic formalism introduced by Freire et. al. to analyze co-operative behavior among

the subunits (motifs) that emerge as result of the evolutive process described here to predict the natural conformation of a protein. One of the simplest structural element exhibiting co-operative folding (and or unfolding) behavior in proteins is the α-helix subunit. This is due to the pattern of hydrogen bonds i.e., between the CO group of the i^{th} amino acid with the NH group of the $(i+4)^{th}$ amino acid. This peculiar pattern inhibits the initiation of a new helix, while enhances the propagation of one which already exists. The co-operative behavior of the α-helix arises from both, the interaction within the subunit, and those with other subunits of the protein. The evaluation of the fitness of a conformer within this framework starts with the assignment of protein secondary structure. Thus, pieces of α-helices are build out of more than five amino acids which comply with standard values of torsional angle ψ and ϕ and of the backbone. We assume in our methodology that α-helices in proteins behave as co-operative folding units and use them as basic structural elements to define higher order folding units. The assessment of goodness using this idea requires also the definition of an imaginary unfolding process of the secondary structure assigned in the first step. The free energy contribution of each of these α-helices in the protein is evaluated using the standard thermodynamic relationship:

$$\Delta G = \Delta H - T\Delta S \tag{1}$$

provided that the enthalpy and entropy changes are known at any temperature.
The enthalpy change associated with the unfolding of an α-helix can be written as:

$$\Delta H = n_h[\Delta H^o + \Delta C_p(T - T^{ref})] \tag{2}$$

where n_h is the number of hydrogen bonds in the helix and T^{ref} is the reference temperature. The difference in heat capacity between the amino acid backbone in the helical and coil conformations has been estimated to be in the order of -7.6 cal/K.mol of amino acid residue. (This is the sum of solvent exposure of the α-carbon in the helical and coil conformations (6.7cal/K.mol) plus the intrinsic heat capacity change associated with the hydrogen bond itself (-14.3cal/K.mol)) [12] . The entropy change associated with the unfolding of an α-helix can be approximated as:

$$\Delta S = n_\alpha \Delta S^o + n_h \Delta C_p \ln(\frac{T}{T^{ref}}) \tag{3}$$

Here n_α is the number of amino acids in the helix and the average entropy per amino acid residue. At T_S^{ref} =112^{o}C the average entropy change for all 20 amino acids is close to 5.5cal/K.mol of amino acid in a α-helix configuration [13]. The co-operative interaction between subunits of protein (helices) can be evaluated from the energetics associated with the change in solvent exposure of the peptide chain upon the imaginary unfolding of any given helix or group of helices within a particular candidate configuration of the protein. The hydrophobic interaction can be defined in terms of the transfer of apolar surfaces from the interior of a protein to water. The energetics of this effect can be calculated from the buried apolar surface area of the number of apolar hydrogen atoms that become exposed to the solvent [14]. For each amino acid residue in the protein, the fraction of inaccessible apolar surface area relative to the total apolar surface area can be determined. The fraction is then multiplied by the number of apolar hydrogen atoms to give the number of buried apolar hydrogen atoms. This calculation is performed for the candidate conformation as well as all the partially folded states

that can be conceived. An example is illustrated for a bundle of 3 helices in Fig.2. Each apolar hydrogen that becomes exposed to water contributes an increment of heat capacity equivalent to 0.398cal/K(mol A^2). The change in heat capacity can then be used to calculate H_{hyd} and ΔS_{hyd} at any temperature using the equations:

$$\Delta H_{hyd} = n_{\alpha h} \Delta C_{p,\alpha h} (T - T_H^{ref}) \tag{4}$$

$$\Delta S_{hyd} = n_{\alpha h} \Delta C_{p,\alpha h} \ln(\frac{T}{T_H^{ref}}) \tag{5}$$

Although the interaction enthalpies and entropies associated with the unfolding of a protein region must take into account effects such as tertiary hydrogen bonds, ionic, and other effects, for simplification of calculations, the interactions considered here are entirely of the hydrophobic type. The free energy contribution of many partially folded states can be thus computed as the sum of the free energies of melting α-helices plus the corresponding free energy of hydrophobic interaction resulting from processes as illustrated in Fig.2

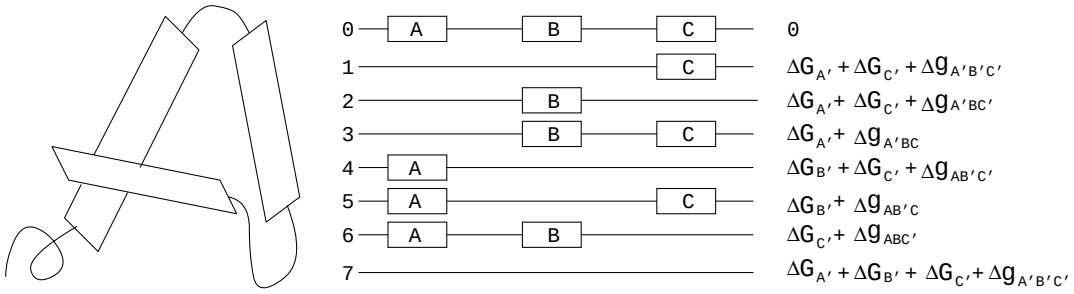

Figure 2: Energetics of a bundle of 3 a-helices. Changes in free energies for melting of helices (upper case G's), and interaction (lower case g's).

We derive scores of thermodynamical stability for a determined protein configuration by computing the changes in free energies for all the conceivable unfolding processes for that configuration. This number equates to 2n, where n is the number of α-helices in the candidate configuration. Here we define the thermodynamic stability coefficient as the sum of the values of free energy change for all the conceivable paths of folding. This sum is our thermodynamical stability score. If the score is large and positive will reflect stability of the conformer, while if it is large and negative, will mean that the structure is more stable in any other conformation. The evaluation of the co-operativity leads to multipoint crossover in the present GA,as co-operative subunits can be interchanged in one crossover operation. Moreover, groups of subunits (motifs) can also be inherited by chromosomes of next generations. The co-operativity analysis routine in our GA is implemented for advanced generations of protein conformers.

2.2 Optimal packing of side chains and the Hybrid GA

Many methods have been proposed to predict the way of packing of the side chains in proteins. C. Lee et. al. [15], for example, apply a simulated annealing algorithm to optimize the side

Figure 3: Polypeptide atoms whose movement is restricted by the algorithm

Figure 4: Parallel Genetic Algorithm on a Network of five transputers

The algorithm is illustrated in Fig.4. The architecture of a transputer, a processor with independent storage memory, consists of a processor endowed with 4 links (hard channels) that allow it to transfer and receive data to or from other four processors of the same type. The algorithm developed here, in its most simple fashion, runs on a net of five transputers. The population of the GA is divided among the processors of the transputer network. The root processor where the master process runs, reads data from the host (a workstation), performs assignment of parameters to the molecule, and transmits data to the slave processes to perform potential function evaluations. The slave processor on the other hand, performs the task of sending data to any of its four neighbor processes, or receive data from any of them. A

generation is built in the network selecting the best individuals from the subpopulation assigned to ea 2 ch processor of the network and sending them into any of the neighbor processors. Similarly, it gets the signal to receive data from any of its neighbors. Each slave process is endowed of routines to copy any individual, to mate the incoming individual from other processor with one of its own individuals.

4 The Addition of Protein-Solvent Interraction

Certainly, our fitness function must include a more elaborate description of solvent-protein interaction. In our previous version of fitness function we included a term based on the calculation of SASA alone. Here, we refine this term by introducing a potential energy term due to solvent-protein interaction as a sum of hydrophobic (E_{hi}) and electrostatic (E_{el}) interactions [8]:

$$E_{se} = E_{hi} + E_{el} \tag{6}$$

Our methodology for the evaluation of solvent-protein interaction is based on a continuum description of the solvent (dielectric-continuum medium). In this model the solvent is characterized by its dielectric constant and SASA. Solvent-protein electrostatic interactions evaluated by means of generalized-Born equation:

$$E_{el} = -\left(1 - \frac{1}{\varepsilon}\right)\left(\sum_{i=1}^{N} \frac{q_i^2}{r_c^i} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{q_i q_j}{d_{ij}}\right) \tag{7}$$

where q_i and q_j are partial electric charges of protein atoms i and j and d_{ij} the distance among them. r_i is the radius of the cavity corresponding to atom i. This equation mixes Born equation for ion solvation and the Coulomb law. We modified the generalized-Born equation in order to account for saturation effects near the protein surface and the non spherical shape of the protein. Our modification of the generalized-Born equation is described in part II of this paper [16]. The hydrophobic interaction term E_{hi} sums up contributions from solvent pressure and solvent-protein van der Waals interaction and is evaluated as a function of protein SASA.

$$H_{hi} = \sum_{i=1}^{N} \sigma_{H,i} A_i \tag{8}$$

The evaluation of this term requires the calibration of the above equation. The surface tension parameters $\sigma_{H,i}$ are obtained by partial least square optimization using a set of solvation enthalpies of small molecules in the solvent under consideration. A detailed presentation of our new solvent-protein interaction model is given in the part II of this paper [16].

5 Results and Conclusions

The hybrid parallel GA was applied to predict the structure of several polypeptides and proteins. We exemplify the performance of the system with the protein CRAMBIN, and we compare the automatically predicted structure with the native crystal structure recorded in the PDB. CRAMBIN is composed of 36 amino acids. The behavior of the energy over five hundred

generations when the hybrid GA is used is ramarkable. The energy passes, at approximately generation 250, from a positive value to a negative value and the algorithm proceeds with the fluctuations peculiar to the evolutionary algorithm. This is the most relevant improvement achieved by the hybridization and parallelization of the original genetic algorithm. A comparison of the best fit individual after the 500^{th} generation obtained by the hybrid GA reveals more accurately the level of evolution of the process. Fig.5 illustrates the conformer obtained by the hybrid GA and the native structure of CRAMBIN as referenced from the PDB. Superposition of both conformers results in a RMS of 6.69 A^o, which although still high, shows the tendency of the molecule to approach that of the crystal structure. A further inspection of the conformers shows patterns of substructures of high similarity in both structures. Comparison of other less similar conformers reveal that these substructural patterns are also being formed and that the evolutionary process as a whole progresses in the direction of more stable conformers, i.e. conformers closer to a minimum.

In conclusion, the hybridicity of the GA proposed in the present work is of the utter most relevance in the evolutionary process for determining the 3D structure of the polypeptide. However, the determination of the effects of only one factor in the whole process of folding is difficult, the improvement of the fitness function in two main aspects such as the co-operativity of structural subunits that begin to form as the generations of better conformers is also a factor that has improved the overall algorithm. This is in fact due to the interaction of fragments that in natural folding processes begin to be stronger than single atom-atom interactions.

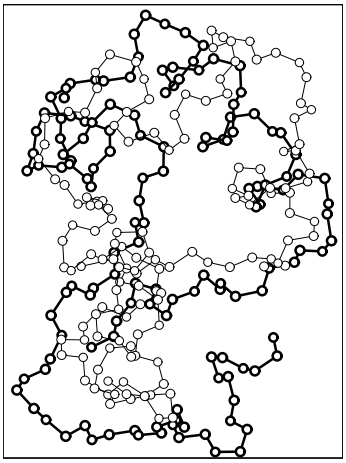

Figure 5: Superposition of the best CRAMBIN conformer obtained by the parallel hybrid GA and the native conformer (bold-line)

Thus our evolutionary algorithm is not only taking into account mere atomic energy relationships, but fragment interactions which reproduce the natural phenomena. Besides having now an improved fitness function, we are also able to evaluate now the thermodynamics of protein solvation (Eqs. 8 and 9 in part II of this paper), which allow us to determine which of the protein conformer will be more stable in solution. In addition, the change in solvation entropy gives us information about solvent reorganization during protein solvation and allows us to assess the strength of hydrophobic interaction of different residues or protein subunits.

The strong point of our new solvent effect model is the availability of first and second-order analytical derivatives of protein SASA. The use of analytical derivatives is expected to lead to a considerably increase in the speed of the optimization part of our hybrid GA. SASA is used in the computation of both the hydrophobic and electrostatic energy terms so the computation of both terms will benefit from the use of analytical derivatives.

References

[1] Gibson K.D., Scheraga H.A.; J. of Comp. Chem. ;11-4, 468-486(1990).

[2] Brugner A.T., Clore G.M., Gronenborn A.M., and Karplus M.; Proc. Natl. Acad. Sci. USA; 83, 3801-3805(1986).

[3] Qian N., Sejnowski T.; J. Mol. Biol.; 202, 865-884(1988).

[4] Del Carpio C.A.; J. Chem. Inf. Comput. Sci.; 36-2, 258-269(1996).

[5] Del Carpio C.A., Sasaki S., Baranyi L., Okada H.; Proc. Genome Informatics Workshop 1995. 130-131(1995).

[6] Del Carpio C., Sasaki S., Proceedings of the XIth International Conference on Computers in Chemical Research and Education. Paris-France 1995

[7] Holland J.; "Adaptation in Natural and Artificial Systems",Ann Arbor: University of Michigan Press (1975)

[8] Goldberg D. E.; "Genetic Algorithms in Search, Optimization, and Machine Learning"; Addi son-Wesley Publ. Inc.; 1989.

[9] Davis L.; "Handbook of Genetic Algorithms"; Van Nostrand Reinhold Inc, New York. 1991.

[10] Koza J.R.; "Genetic Programming"; Massachusetts Institute of Technology; 1993.

[11] Freire E., and Murphy P. K.; J. Mol. Bio.; 222, 687-698(1991).

[12] Murphy K.Pl, and Gill S.J.; Thermochim. Acta, 172, 11-20(1990).

[13] Murphy K.P. , Privalov P.L, Gill S.J.; Science 222, 559-561 (1990).

[14] Lee B., Richards F.M.; J. Mol. Biol 55, 379-400(1971).

[15] Lee C., Subbiah S.; J. Mol. Biol.; 217, 373- 388 (1991)

[16] Allinger N.L, Tribble M.T., Miller D.H.; J. Am. Chem. Soc.; 1971; 93(1637)

[17] Gogonea V., Gogonea B.C., Del Carpio C. A.;Proc. Genome Informatics 1996.